Mechanistic insights from combining genomics with metabolomics

Fotios Drenos

MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol,

Oakfield House, Oakfield Grove, Bristol, BS8 2BN

Address for correspondence:

MRC Integrative Epidemiology Unit at the University of Bristol

University of Bristol

Oakfield House

Oakfield Grove

Bristol BS8 2BN

E-mail: f.drenos@bristol.ac.uk

Abstract

Purpose of review: Metabolomics directly measure substrates and products of biological processes and pathways. Based on instrumentation and throughput advances, the use of metabolomics has only recently become feasible at the population level. This has led to an intense interest in using the new information in combination to genomics, and other –omics technologies, to give biological context to the rapidly accumulating associations between genes and diseases or their risk factors.

Recent findings: The use of metabolomics-genomics associations for the metabolic characterisation of genes of interest have confirmed known pathways and permitted the identification of new ones. These include the unknown metabolite X12063 linking satins to myopathies, the role of glycerophospholipids in cholesterol metabolism, the structure of lipoprotein (a), the LPL-independent effect of *APOC3* and the role of branched chain amino acids in the antagonistic co-regulation of high density lipoproteins and triglyceride levels.

Summary: The findings reviewed illustrate the importance of integrating metabolomics and genomics for the greater understanding of biological mechanisms. The limitations of the current approaches are also discussed together with what will be required in order to make the most of the current multi-omics data available.

Keywords: Metabolomics, genomics, multi-omics, mGWAS

Introduction

Metabolomics is the study of the quantitative complement of small molecules in biological systems. The metabolic measures obtained are mostly organic compounds involved in the biochemical reactions of the organism and represent the final stage of the flow of information from the genome to the transcriptome and then to the proteome before it reaches the biological phenotype [1]. Unlike the genes and proteins though, that can be subjected to epigenetic or post-translational changes, the metabolites provide a direct signature of biochemical activity and a snapshot of the underlying state of the organism or specific biological system sampled [2]. In the case of lipids, lipidomics, a subfield of metabolomics, can measure not only the larger groupings of lipids, commonly assessed in the clinic,, such as high density lipoproteins (HDL), low density lipoproteins (LDL), triglycerides (TG) and total cholesterol, but also their specific subclasses both in terms of concentration and composition, as well as a large number of other lipid compounds involved in numerous physiological processes or biological structures. The measurements obtained from the metabolomics approaches can be used to study the relationships between the metabolites or used to better characterise a complex biological mechanism.

Genomics study the function of the genome and its role in the manifestation of the organism's characteristics or risk of disease. The genome can be divided in a protein coding part, known as the exome, and a non-coding part. Changes in both have been associated with a number of diseases or phenotypic changes in humans and other organisms. Very large efforts have focused on cataloguing all functional variation present in the genome, either in coding [3, 4] or non-coding regions [5]. Despite these efforts and the large number of data currently available, in the majority of cases, we are still unsure on how a DNA change results in changes in the observed phenotype.

Since metabolomics represent phenotyping at the molecular level that extends closer to the biological processes taking place in the organism, which are mostly coded by the genome, the

integration of the two levels of information can provide the opportunity to better study biological mechanisms and how they result in a final phenotypic change or disease.

Metabolic traits genome wide association

The most straightforward method to combine metabolomics with genomics is the use of a metabolic traits genome wide association study (mGWAS). Unlike studies trying to uncover associations with a risk factor or disease endpoint, mGWAS usually identify associations that point directly to specific underlying biological mechanisms [6]. Commonly multiple metabolites are tested in parallel mGWAS, either based on a preselection of a small number of metabolites or using the whole set available. Trying to identify endogenous molecules and genetic factors affecting lean mass, Korostishevsky et al [7] performed an association analysis between a lean muscle phenotype and an array of metabolomics measures. They then selected the top 3, explaining 11% of the variation of the trait, to run the mGWAS. One of the three metabolites used, X12063, was found to be associated with polymorphisms in the *CYP3A5* and *SLCO1B1* genes. Both loci have previously been associated with statins uptake and efficacy and statins are known for their muscle related side-effects [8]. Unfortunately, X12063 is currently an unknown metabolite but based on its associations with the two genes we will hopefully soon be able to identify it and further understand the link between statins and myopathies.

Characterising the metabolic impact of genes

A more common use of mGWAS is to provide a functional characterisation of specific genes and allow us to better understand the relevant pathways. This is closer to what is commonly described as a phenome scan, in this case metabolome scan, where a locus of interest is tested for association with multiple phenotypes, here metabolites or metabolomics measures. In an illustrative example,

using results from 129 mGWAS, Draisma et al [9], found 31 loci associated with 85 different metabolites. The *FADS1-3* gene cluster showed the highest number of associations, with 12 SNPs in the locus associated with 47 of the metabolites tested, all of which were glycerophospholipids. SNPs in the locus have also been associated with LDL, HDL, total cholesterol and TG levels in plasma [10, 11], as well as polygenic dyslipidemia [12]. The function of the genes in the cluster is to introduce double bonds during the desaturation of fatty acids in a well characterised pathway. The glycerophospholipids measured were both up- and down-stream of the enzymatic action of the gene products and their changes reflected the known metabolic pathway steps [13]. According to these results, the balance between glycerophospholipids and the availability of polyunsaturated long-chain fatty acids with four and more double bonds has the potential to alter the levels of cholesterol, TG and associated lipoproteins. Furthermore, the close agreement between what is known for the function of the FADS1-3 cluster of genes from biochemical studies and the metabolomics results obtained from population samples, through the combination of genetic and metabolomics data, provides a perfect illustration of the power of this method to gain insights into biological processes *in vivo* which were otherwise hidden from observational epidemiology at this level.

Insights in Lipoprotein (a) composition

In another example of mGWAS use, Kettunen et al [14] tested the genome wide associations of 123 metabolic measures focusing on the characterisation of the *LPA* gene. *LPA* codes for Lipoprotein (a) (Lp(a)) a well-established coronary heart disease (CHD) risk factor [15]. Elevated levels of Lp(a) are believed to increase the risk of cardiovascular disease through the inhibition of fibrinolysis and the promotion of coagulation [15]. Except the already known associations with LDL, total cholesterol and TG, the analysis of the metabolomics measures revealed strong associations with the diameter and concentration levels of the larger very low density lipoprotein (VLDL) particles [14, 16]. Using a gene score, accounting for 45% of the Lp(a) variance in the sample used, to summarise the total effect of the associated LPA polymorphisms on Lp(a) protein, the authors repeated the analysis of the

metabolomics measures obtaining similar results, with the diameter of the VLDL particles being the top signal. Comparing the pattern and estimates of the associations of the Lp(a) protein levels to the metabolites and the associations of the LPA polymorphisms on the same metabolites suggested a causal effect of Lp(a) on the relevant VLDL measures. Based on these results, the authors suggested that the apoB-containing lipoprotein particle, which covalently binds to apo(a) for the formation of Lp(a), may also be a poorly lipidated VLDL-type of particle, challenging the idea that Lp(a) is just an apo(a) component added to LDL particles [14].

The function of APOC3 and its LPL independent role

The use of metabolomics to characterise rare genetic variants altering the structure or function of the coded protein, provides us with an even better direct view of the relevant biological mechanisms operating in the organism. The APOC3 protein is involved in a number of intra- and extra- cellular mechanisms including the production and clearance of triglyceride rich lipoproteins. Polymorphisms in the *APOC3* gene locus have been previously robustly associated with levels of plasma TG, HDL and VLDL as well as risk of CHD [17-22]. A rare loss of function mutations in the gene, rs138326449 (IVS2+1G>A), significantly lowers the risk of CHD by approximately 40% [18, 20]. The same mutations was tested against 225 metabolomics measures including the size and composition of 14 lipoprotein subclasses in 13,285 participants from two European population cohorts [23]. In addition to the previously reported associations, the variant was also strongly associated with VLDL and HDL composition measures, other cholesterol measures and fatty acids. The effect of APOC3 on lipids is believed to operate through the inhibition of triglyceride rich lipoproteins hydrolysis by LPL which in turn lowers their uptake from hepatocytes. Under this model, the decrease in VLDL and increase of HDL concentrations were due to the loss of function mutation permitting a faster LPL mediated rate of hydrolysis. To test this hypothesis, the common *LPL* single nucleotide polymorphism (SNP), and leading GWAS association SNP in the locus, rs12678919 was tested for association with the 225 measures. The results confirmed that the two polymorphisms, from *APOC3* and *LPL*, had the same

general pattern of associations, providing additional support for the standard model. Mathematical modelling of the individual SNP-measures associations though, revealed that this is not true for all of the lipid traits tested. The triglyceride content of very large and medium VLDL particles was lower than expected if the effect was solely through LPL inhibition. VLDL is produced in the hepatocytes through a complex process, not yet fully understood, where APOB-100 is combined with triglyceride and other apolipoproteins and cholesteryl esters [24, 25]. In cell cultures and model organisms under insulin resistance or hypertriglyceridemia, APOC3 facilitates the binding of triglycerides to pre-VLDL molecules. Specific changes in the APOC3 protein affect the composition of larger VLDL particles but not the less TG rich smaller VLDL [25]. Although the metabolomics associations in this case refer to changes in the plasma VLDL composition, it is evident that they reflect the end-product of an intracellular process affecting the assembly and secretion of the larger VLDL particles. The results also suggest that the role of APOC3 in the cell, although noticeable under specific conditions during which the liver is trying to remove stored TG, is continuously active. Though, the contribution of the LPL-independent pathway to the overall levels of VLDL and TG is rather limited under normal conditions in comparison to the LPL-dependent pathway. The same intra-cellular APOC3 mechanism might also contribute to the effect of the *APOC3* antisense oligonucleotide inhibitor ISIS 304801 which showed promising results for the lowering of TG levels in severe hypertriglyceridemia patients and familial chylomicronemia syndrome sufferers [26, 27].

Metabolomics in multi-omics approaches

The greatest promise of metabolomics lies in taking into account the organisation of the metabolites in defined pathways or groups and combining this information with genomics or other –omics technologies. Various methods are currently in use on metabolomics that consider metabolites, not in isolation, but organised in pathways. These can be based on previous information from the available pathway databases (for example KEGG [28]), data that can be used to infer the functional

relationship of the metabolites, or a combination of known information and new results [29]. The pathway information can then be used to test for association with a gene, or group of genes, describing the biological mechanism of interest. An example of such as approach can be found in Ried et al [30] which used a phenotype set enrichment analysis to test for over-representation of sets of metabolites at genes. Their results suggested that genes of the cytochrome P450 family 4 are involved in reactions that include glycerolipids and fatty acids as substrates, haeme as a cofactor and carnitines as related compounds. Even more information on the underlying molecular mechanism can be obtained from metabolomics and genomics data by incorporating additional –omics results. Recently, Bartel et al [31] combined whole blood transcriptomics with metabolomics and used the genomic information to test the causal effects between transcripts and metabolites. Further integration with external database information allowed the identification of systematic signatures of lipid, energy and amino acid metabolic reactions. Using the derived functional categories in a network representation of the system, revealed pathways that are activated in a coordinated manner. Finally the measured metabolites, transcripts, metabolic pathways and gene annotation terms were used to identify novel associations between the molecular mechanisms inferred and the common CHD risk factors of HDL, LDL and TG. An antagonistic pattern of associations between HDL and TG was observed for parts of the model where co-regulation was evident. Interestingly, a regulatory mechanism involving the sterol regulatory element-binding genes was found to link HDL, TG and branched chain amino acids at a transcriptional level. A negative association between HDL and the "gamma-glutamyl metabolism" pathway was also found, which was similar to what was observed through the association of a *GLUL* gene polymorphism, encoding for the glutamine-synthase enzyme, to HDL [32]. The later though was seen only in those with type-2-diabetes, so the proposed mechanism might actually be more complex than we think and might involve other metabolites and higher order interactions.

Present challenges and limitations

The integration of metabolomics with other –omics data, and most importantly genomics, is key to the fulfilment of the metabolomics promise for precision medicine [33]. The data allowing such multilevel approaches, have only recently started to become available in a scale that permits well powered epidemiological studies with independent replication. Though, a number of limitations are starting to also become obvious. There is no reliable estimate on how big the human metabolome is. We are currently able to measure and provide quantification for hundreds of known metabolites and relative quantification for thousands known and unknown measures from biological samples such as blood and urine [33]. The Human Metabolome Database (http://www.hmdb.ca/) has catalogued 42,003 metabolites with just 5,701 linked to protein sequences and more are expected to exist. Currently, there is no single method that can provide a comprehensive measurement of all the known metabolites. The identification of the large number of unnamed metabolites is still an issue which requires substantial effort. Methods to predict what a unnamed metabolite is are available, but they still require laboratory verification which is not easy [34]. There is also significant fragmentation in the laboratory and computational methods used to obtain, interpret and quantify raw metabolomics data. The majority of these are protected behind intellectual property rights [35] that make the standardisation difficult. A recent effort to compare 43 common metabolites from two platforms in the same individuals found that their mean correlation was just 0.44, though some showed very high agreement and some almost none [36]. All this issues are expected to be resolved in the next few years as the metabolomics technology matures and a standard emerges or prevails.

Conclusions

Despite the limitations, the integration of metabolomics with genomics has already provided a wealth of new information on the function of genes that have been previously associated with important diseases and the pathways through which they exert their effects. Use of additional –

omics to further enriched the available information will be the key for the complete understanding of the biological mechanisms maintaining human health. This multi-omics view though will also require sustained cross-field collaborations and a new generation of scientists that can communicate across fields as diverse as chemistry, epidemiology and computer science. Only when the integration of the data is supported by an equal integration of ideas we will start to translate the current abundance of data to patients benefit.

Key points

- Metabolomics measure the molecular phenotype of the biological processes taking place in the organism
- Parallel metabolic trait genome wide association analyses are currently used to describe the function mechanism of genes and elucidate the corresponding pathways.
- Multi-omics approaches make use of metabolomics in combination with genomics and transcriptomics to present a system wide picture of the relationships between metabolites.
- A number of challenges, mainly around the part of the metabolome captured and the compatibility of the measures obtained from different platforms, remain in the field and affect the use of the data.

Acknowledgements

None

Financial support and sponsorship

Fotios Drenos is supported by the UK Medical Research Council (MC_UU_12013/1-9)

Conflicts of interest

The author does not have any conflicts of interest in the subject area.

References

1. Dunn WB, Broadhurst DI, Atherton HJ, et al. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews*. 2011;40:387-426.
2. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omic triology. *Nature reviews Molecular cell biology*. 2012;13:263-269.
3. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285-291.
4. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*. 2012;22:1760-1774.
5. ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57-74.
6. Dharuri H, Demirkan A, van Klinken JB, et al. Genetics of the human metabolome, what is next? *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2014;1842:1923-1931.
7. Korostishevsky M, Steves CJ, Malkin I, et al. Genomics and metabolomics of muscular mass in a community-based sample of UK females. *Eur J Hum Genet*. 2016;24:277-283.
8. Thompson PD, Panza G, Zaleski A, Taylor B. Statin-Associated Side Effects. *J Am Coll Cardiol*. 2016;67:2395-2410.
9. Draisma HHM, Pool R, Kobl M, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature Communications*. 2015;6:7208.
10. Asselbergs Folkert W, Guo Y, van Iperen Erik P, et al. Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci. *American Journal of Human Genetics*. 2012;91:823-838.
11. Willer CJ, Sanna S, Jackson AU, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*. 2008;40:161-169.
12. Kathiresan S, Willer CJ, Peloso G, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics*. 2009;41:56-65.
13. Gieger C, Geistlinger L, Altmaier E, et al. Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. *PLoS Genet*. 2008;4:e1000282.
14. ******Kettunen J, Demirkan A, Würtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications*. 2016;7:11122. **Recent mGWAS for 123 lipids related metabolomics measures that provide full summary data for all mGWAS considered**.
15. Nordestgaard BG, Chapman MJ, Ray K, et al. Lipoprotein(a) as a cardiovascular risk factor: current status. *European Heart Journal*. 2010;31:2844-2853.
16. The Emerging Risk Factors C. Lipoprotein(a) Concentration and the Risk of Coronary Heart Disease, Stroke, and Nonvascular Mortality. *JAMA : the journal of the American Medical Association*. 2009;302:412-423.
17. Asselbergs Folkert W, Guo Y, van Iperen Erik PA, et al. Large-Scale Gene-Centric Meta-analysis across 32 Studies Identifies Multiple Lipid Loci. *Am J Hum Genet*. 2012;91:823-838.
18. Crosby J, Peloso GM, Auer PL, et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med*. 2014;371:22-31.
19. Global Lipids Genetics C. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*. 2013;45:1274-1283.
20. Jorgensen AB, Frikke-Schmidt R, Nordestgaard BG, Tybjaerg-Hansen A. Loss-of-function mutations in APOC3 and risk of ischemic vascular disease. *N Engl J Med*. 2014;371:32-41.
21. Tachmazidou I, Dedoussis G, Southam L, et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nature Communications*. 2013;4.
22. Timpson NJ, Walter K, Min JL, et al. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Communications*. 2014;5:4871.

23. *****Drenos F, Davey Smith G, Ala-Korpela M, et al. Metabolic Characterization of a Rare Genetic Variation within APOC3 and its Lipoprotein Lipase Independent Effects. *Circulation: Cardiovascular Genetics*. 2016.
Example of the use of mathematical modelling for the distinction between the LPL dependent and LPL independent effect of APOC3
24. Sundaram M, Yao Z. Recent progress in understanding protein and lipid factors affecting hepatic VLDL assembly and secretion. *Nutrition & Metabolism*. 2010;7:1-17.
25. Yao Z, Wang Y. Apolipoprotein C-III and hepatic triglyceride-rich lipoprotein production. *Curr Opin Lipidol*. 2012;23:206-212.
26. Gaudet D, Alexander VJ, Baker BF, et al. Antisense Inhibition of Apolipoprotein C-III in Patients with Hypertriglyceridemia. *N Engl J Med*. 2015;373:438-447.
27. Gaudet D, Brisson D, Tremblay K, et al. Targeting APOC3 in the Familial Chylomicronemia Syndrome. *N Engl J Med*. 2014;371:2200-2206.
28. Kanehisa M, Goto S, Sato Y, et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*. 2014;42:D199-D205.
29. *****Ren S, Hinzman AA, Kang EL, et al. Computational and statistical analysis of metabolomics data. *Metabolomics*. 2015;11:1492-1513.
Very good review of statistical and computational meathods in use in metabolomics
30. Ried JS, Shin S-Y, Krumsiek J, et al. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Human Molecular Genetics*. 2014;23:5847-5857.
31. *****Bartel J, Krumsiek J, Schramm K, et al. The Human Blood Metabolome-Transcriptome Interface. *PLoS Genet*. 2015;11:e1005274.
A multi-omics example in practise showing how graphical representation of metabolomics can improve our understanding of the underlying mechanisms
32. Beaney KE, Cooper JA, McLachlan S, et al. Variant rs10911021 that associates with coronary heart disease in type 2 diabetes, is associated with lower concentrations of circulating HDL cholesterol and large HDL particles but not with amino acids. *Cardiovascular Diabetology*. 2016;15:115.
33. Beger RD, Dunn W, Schmidt MA, et al. Metabolomics enables precision medicine: "A White Paper, Community Perspective". *Metabolomics*. 2016;12:149.
34. Krumsiek J, Suhre K, Evans AM, et al. Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genet*. 2012;8:e1003005.
35. Ala-Korpela M, Davey Smith G. Metabolic profiling–multitude of technologies with great research potential, but (when) will translation emerge? *International Journal of Epidemiology*. 2016.
36. Yet I, Menni C, Shin S-Y, et al. Genetic Influences on Metabolite Levels: A Comparison across Metabolomic Platforms. *PLoS ONE*. 2016;11:e0153672.